Optimal selection of individuals for repeated covariate measurements in follow-up studies

Jaakko Reinikainen[1,*], Juha Karvanen[1] and Hanna Tolonen[2]

[1] Department of Mathematics and Statistics,
University of Jyväskylä, Jyväskylä, Finland

[2] Department of Chronic Disease Prevention,
National Institute for Health and Welfare, Helsinki, Finland

*Corresponding author:
Jaakko Reinikainen, Department of Mathematics and Statistics,
P.O. Box 35 (MaD), FI-40014 University of Jyväskylä, Finland
Email: jaakko.o.reinikainen@jyu.fi, tel.: +358 440 366 896

January 23, 2014

Abstract

Repeated covariate measurements bring important information on the time-varying risk factors in long epidemiological follow-up studies. However, due to budget limitations, it may be possible to carry out the repeated measurements only for a subset of the cohort. We study cost-efficient alternatives for the simple random sampling in the selection of the individuals to be remeasured. The proposed selection criteria are based on forms of the D-optimality. The selection methods are compared in simulation studies and illustrated with the data from the East-West study carried out in Finland from 1959 to 1999. The results indicate that cost savings can be achieved if the selection is focused on the individuals with high expected risk of the event and, on the other hand, on those with extreme covariate values in the previous measurements.

Keywords: optimal design, data collection, repeated measurements, follow-up study, missing covariate data

1 Introduction

Many epidemiological follow-up studies include covariates, such as blood pressure, cholesterol and weight, that may vary over the time. If only the baseline measurement of these covariates is used, the analysis may suffer from the regression dilution problem,[1,2] i.e. the measurement made a long time ago does not anymore predict the disease risk. To avoid this problem, we need to carry out repeated measurements on the time-varying covariates. Conducting these measurements in a large population sample is expensive, and due to budget limitations, we may not be able to select the entire cohort for the new measurement, but only a subset of it. In this article, we study how this subset should be selected in order to estimate the effect of the covariate on survival time as accurately and precisely as possible. We propose the subset to be selected so that a Fisher information based optimality criterion will be optimized.

The planning of cost-efficient study designs has led to the development of multi-stage (or sequential) designs. In multi-stage studies, the next stage is constructed on the basis of the information obtained from earlier stages under given budget limitations. Multi-stage designs allow us to allocate the sample size optimally between the stages.[3,4] Even further, optimality criteria developed originally for design of experiments[5,6] can also be applied in observational multi-stage studies. Karvanen et al. explore optimal ways to select a small subset of individuals for expensive genotyping in follow-up studies.[7] Mehtälä et al. consider optimal designs for the measurement times of a binary continuous-time Markov process.[8]

The research question outlined above is explored here in a simplified setup with one time-varying covariate and two measurement times using simulated and real data. Several authors have considered general approaches for the joint modeling of survival data and longitudinal covariate data.[9–12] The real data come from the Finnish cohorts of the Seven Countries Study[13] carried out in Finland from 1959 to 1999. The objective of The Seven Countries Study was to investigate variation in cardiovascular disease and related risk factors levels. In our example, diastolic blood pressure is the time-varying covariate of interest, and the survival time is considered to be the age at the time of death. With these data, we compare the results corresponding to different selection methods to the situation where every individual is selected

for the second measurement.

As we are selecting only a subset of individuals for the new measurement, the majority of individuals does not have this measurement, and thus the handling of missing data plays an important role in the modeling. We study here two alternative ways to deal with the missing data: multiple imputation and a likelihood-based approach with numerical integration.[14]

The underlying model is described in Section 2 and the selection criteria are presented in detail in Section 3. In Section 4, we discuss an algorithm, which is used in finding optimal or nearly optimal designs with respect to the chosen criterion. Statistical analysis of data collected during the whole follow-up, is described in Section 5, with an emphasis on the handling of missing data. Simulation studies comparing different selection methods are presented in Section 6, followed with results obtained using the real data in Section 7. Finally, Section 8 concludes the paper.

2 Survival model

We consider a follow-up study with a predetermined length, where all individuals have a baseline measurement of a single time-varying covariate. The outcome variable is survival time, which is censored at the end of the follow-up. The second measurement of the covariate will be taken halfway through the follow-up. Assume, that we cannot afford to remeasure the entire cohort, and therefore have to select a subset of individuals for the second measurement. Here, we study cost-efficient alternatives for the simple random sampling in this selection, which is conducted just before the time of the second measurement. The interest lies in the utilization of the baseline measurement information. In addition, the age of an individual may also affect the selection as we are operating with the age as the time axis in our Weibull proportional hazards model.

We start by specifying the survival model, which we need later in defining the selection criteria. Let $Y_1 = (t_1, \delta_1)$ denote the survival information at the time of the second measurement, where continuously measured time t_1 is censored if the individual is still alive. The status indicator gives $\delta_1 = 1$ for the event and $\delta_1 = 0$ for censoring. For the second part of the follow-up we use similar notation $Y_2 = (t_2, \delta_2)$.

Under the proportional hazards model with a time-varying covariate $x(t)$ the hazard function has the form

$$\lambda(t|x(t)) = \lambda_0(t)e^{\beta x(t)},$$

where parameter β describes the relation between the covariate and survival

time. We continue by assuming the Weibull distribution for the survival times. The baseline hazard function is parameterized with shape parameter a and scale parameter b as

$$\lambda_0(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1}.$$

Now, assume a piecewise constant covariate for an individual j:

$$x_j(t) = \begin{cases} x_{0j}, & t \in [t_{0j}, t_{1j}) \\ x_{1j}, & t \in [t_{1j}, t_{2j}), \end{cases}$$

when the hazard can be written as

$$\lambda(t_{1j}|x_{0j}) = \frac{a}{b}\left(\frac{t_{1j}}{b}\right)^{a-1} e^{\beta x_{0j}}$$

for the first part of the follow-up and as

$$\lambda(t_{2j}|x_{1j}) = \frac{a}{b}\left(\frac{t_{2j}}{b}\right)^{a-1} e^{\beta x_{1j}}$$

for the second part of the follow-up. The whole data for N individuals are denoted by (X_0, X_1, Y_1, Y_2).

In general form, the survival function and the density function are

$$\begin{aligned} S(t|x(t)) &= S_0(t)^{\exp(\beta x(t))} \text{ and} \\ f(t|x(t)) &= \lambda(t|x(t))S(t|x(t)), \end{aligned}$$

where $S_0(t)$ is the baseline survival function. Because we are operating with age as the time axis and individuals enter the follow-up at different ages, we have to deal with truncated distributions. Denote the survival time (age) at entering the follow-up by t_0. The likelihood function of parameters $\theta = (\beta, a, b)$ for an individual j has the form

$$L_j(\theta) = \left(\frac{f(t_{1j}|x_{0j})}{S(t_{0j}|x_{0j})}\right)^{\delta_{1j}} \left(\frac{S(t_{1j}|x_{0j})}{S(t_{0j}|x_{0j})}\right)^{1-\delta_{1j}} \left(\frac{f(t_{2j}|x_{1j})}{S(t_{1j}|x_{1j})}\right)^{\delta_{2j}} \left(\frac{S(t_{2j}|x_{1j})}{S(t_{1j}|x_{1j})}\right)^{1-\delta_{2j}},$$

where the first two factors of the product correspond to information from the first half of the follow-up and the last two factors correspond to information from the second half of the follow-up. The survival functions in denominators are needed to scale distributions, because we do not assume the follow-up to start from the time origin. An individual has contribution to the part of the likelihood concerning the second part of the follow-up only if he or she has not had an event before the second measurement.

3 Selection criteria

Our main question is, how to select a subset of individuals for the second measurement if only n individuals can be selected. We would like to select a subset which allows parameter β to be estimated as accurately and precisely as possible. Applying the principles of optimal experimental design, the selection criteria are based on the Fisher information matrix of parameters $\theta = (\beta, a, b)$

$$I_{X,Y}(\theta) = -E_\theta \left(\frac{\partial^2 \log p(X_0, X_1, Y_1, Y_2)}{\partial \theta^2} \right),$$

where at the time of the selection X_0 and Y_1 are observed and X_1 and Y_2 are unknown. By factorizing the logarithmic joint distribution of X_0, X_1, Y_1 and Y_2 with the assumption $p(Y_2|X_0, X_1, Y_1) = p(Y_2|X_1, Y_1)$ it follows

$$\log p(X_0, X_1, Y_1, Y_2) = \log p(X_0) + \log p(Y_1|X_0) + \log p(X_1|X_0, Y_1) + \log p(Y_2|X_1, Y_1)$$

and the information matrix becomes

$$\begin{aligned} I_{X,Y}(\theta) &= -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p(X_0) + \frac{\partial^2}{\partial \theta^2} \log p(Y_1|X_0) \right. \\ &\quad \left. + E_\theta \left(\frac{\partial^2}{\partial \theta^2} \log p(X_1|X_0, Y_1) \middle| X_0, Y_1 \right) + E_\theta \left(\frac{\partial^2}{\partial \theta^2} \log p(Y_2|X_1, Y_1) \middle| X_0, Y_1 \right) \right] \\ &= E_\theta \left(-\frac{\partial^2}{\partial \theta^2} \log p(Y_1|X_0) \right) + E_\theta \left[E_\theta \left(-\frac{\partial^2}{\partial \theta^2} \log p(Y_2|X_1, Y_1) \middle| X_0, Y_1 \right) \right] \\ &= I_{Y_1|X_0}(\theta) + E_\theta(I_{Y_2|X_1,Y_1}(\theta)). \end{aligned} \tag{1}$$

Above, the terms $\log p(Y_1|X_0)$ and $\log p(X_1|X_0, Y_1)$ do not include parameters θ and cancel out. The likelihood contribution to $p(Y_2|X_1, Y_1)$ comes only from the individuals, who have not had an event before the time of the second measurement, i.e. individuals with $\delta_1 = 0$.

In practice, we replace the first term of (1) by observed information $J_{Y_1|X_0}(\theta)$, and get a mixture of the observed and expected information matrices, which has the elements

$$\begin{aligned} \Psi_{X,Y}(\theta)_{i,j} &= J_{Y_1|X_0}(\theta)_{i,j} + E_\theta(I_{Y_2|X_1,Y_1}(\theta))_{i,j} \\ &= -\sum_{k=1}^{N} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(Y_{1k} = y_{1k}|X_{0k} = x_{0k}) \right] \\ &\quad -\sum_{k=1}^{n} \left[E_\theta \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(Y_{2k}|X_{1k}, Y_{1k} = y_{1k}) \middle| X_{1k}, Y_{1k} = y_{1k} \right) \right], \end{aligned}$$

where the first term consists of the information from the first period of the follow-up from all individuals $k = 1, \ldots, N$. The second term includes the information from the second measurement, thus the sum is only over the selected subset of individuals $k = 1, \ldots, n$. As the selection of individuals is carried out just before the time of the second measurement, the variables X_1 and Y_2 are not observed for anyone and the expectation above can be calculated by Monte Carlo integration.[15]

The two selection criteria we will apply, are well known criteria from the theory of optimal designs. Methods we will apply in the selection are based on forms of the D-criterion. The D-optimal design is obtained by maximizing the determinant of the information matrix $\det(\Psi_{X,Y}(\theta))$. This is equivalent to minimizing $\det(\Psi_{X,Y}(\theta)^{-1})$.

The second method we are using is the D_β-criterion. D_β-optimal design is obtained by minimizing the D_β-criterion, which we define as a diagonal element corresponding to β of the 'covariance matrix' $\Psi_{X,Y}(\theta)^{-1}$. In other words, we want to minimize $\mathrm{Var}(\hat{\beta})$. This criterion is a special case of D_s-optimality, where we are interested in a subset of s parameters.[5] In our case this subset consists of the parameter β.

The calculation of optimality criteria requires the second order partial derivatives of $\log p(y_1|x_0)$ and $\log p(y_2|x_1, y_1)$. First, considering $\log p(y_1|x_0)$, we calculate the second order partial derivatives of

$$\log p(y_1|x_0) = \log \left[\left(\frac{f(t_1|x_0)}{S(t_0|x_0)} \right)^{\delta_1} \left(\frac{S(t_1|x_0)}{S(t_0|x_0)} \right)^{1-\delta_1} \right],$$

which are

$$\begin{aligned}
\frac{\partial^2}{\partial \beta^2} \log p(y_1|x_0) &= x_0^2 e^{\beta x_0} \log \left(\frac{S_0(t_1)}{S_0(t_0)} \right), \\
\frac{\partial^2}{\partial a^2} \log p(y_1|x_0) &= -\delta_1 a^{-2} - e^{\beta x_0} \left[\log \left(\frac{t_1}{b} \right) \right]^2 \left(\frac{t_1}{b} \right)^a + e^{\beta x_0} \left[\log \left(\frac{t_0}{b} \right) \right]^2 \left(\frac{t_0}{b} \right)^a, \\
\frac{\partial^2}{\partial b^2} \log p(y_1|x_0) &= \delta_1 \frac{a}{b^2} + e^{\beta x_0} a(-a-1) b^{-a-2} (t_1^a - t_0^a), \\
\frac{\partial^2}{\partial \beta \partial a} \log p(y_1|x_0) &= -x_0 e^{\beta x_0} \left[\left(\frac{t_1}{b} \right)^a \log \left(\frac{t_1}{b} \right) - \left(\frac{t_0}{b} \right)^a \log \left(\frac{t_0}{b} \right) \right], \\
\frac{\partial^2}{\partial \beta \partial b} \log p(y_1|x_0) &= x_0 e^{\beta x_0} a b^{-a-1} (t_1^a - t_0^a) \text{ and} \\
\frac{\partial^2}{\partial a \partial b} \log p(y_1|x_0) &= -\delta_1 \frac{1}{b} + e^{\beta x_0} b^{-a-1} \left[t_1^a a \log \left(\frac{t_1}{b} \right) + t_1^a - t_0^a a \log \left(\frac{t_0}{b} \right) - t_0^a \right].
\end{aligned}$$

For $\log p(y_2|x_1, y_1)$ corresponding formulae are obtained by replacing $f(t_1|x_0)$, $S(t_1|x_0)$ and $S(t_0|x_0)$ by $f(t_2|x_1), S(t_2|x_1)$ and $S(t_1|x_1)$, respectively.

From these formulae, it cannot be directly seen what kind of individuals would be included in the D_β-optimal or D-optimal subsets. Intuitively, we could expect that individuals with extreme covariate values or with high risk of the event would be important for the estimation. In a case of first-order linear regression models extreme selection of covariate values is optimal,[16] but this does not hold e.g. for quadratic models or binary responses. Extreme selection means selecting individuals with highest and lowest covariate values. Although this method could be applied also in our case, there is no guarantee that it would be optimal in any sense, because we consider a survival model instead of a linear regression model. On the other hand, in survival models, events contain more information than censorings. Intuitively, this means that individuals, who are likely to have an event during the follow-up should be selected. Results with simulated and real data in Sections 6 and 7 show, how these intuitive principles will be balanced according to D_β-optimal and D-optimal selections.

4 Finding optimal designs

As the optimal designs of nonlinear models depends on the parameters, we need initial estimates for parameters a, b and β, to use optimality criteria in practice. These can be obtained from the data already collected during the follow-up before the time of second measurement and/or from previous studies. For study cohorts of reasonable size, it is not computationally possible to explore all possible subsets of individuals to be selected for the new measurement but heuristic methods are needed. We use a greedy method,[17] where the individuals are selected sequentially one by one to find the optimal subset. This method is also known as the sequential search[18] and is a simplified special case of the Fedorov-Wynn algorithm[19]. The jth individual is selected so that the selection criterion is optimized on the condition that $j-1$ individuals have already been selected.

Denote the set of individuals already selected by S. Using D_β-optimality, the next individual $j \notin S$ is selected so that $\mathrm{Var}(\hat{\beta})$ obtained from the appropriate diagonal element of

$$\left(\sum_{i \in S} \Psi_{x_i, y_i}(\hat{\theta}) + \Psi_{x_j, y_j}(\hat{\theta}) \right)^{-1}$$

is minimized. To find the D-optimal design, the selection is carried out so

that

$$D = \det\left(\sum_{i \in S} \Psi_{x_i,y_i}(\hat{\theta}) + \Psi_{x_j,y_j}(\hat{\theta})\right)$$

is maximized. In a case in which there are more than one individual that optimizes the criterion at issue, the selection between them is done randomly. The selected individual is added to the set S and the procedure continues until the set S has reached the predetermined size.

In general, the selection problem is NP-hard and the greedy method produces only a suboptimal solution. However, the empirical results[7] indicate that the gain from more complicated heuristics may not be large in this kind of design problem.

5 Statistical analysis

5.1 Likelihood-based approach with numerical integration

After the follow-up study we have data with a large amount of missing data, as only a subset is selected for the second measurement. When that subset is selected using D_β-optimality or D-optimality, the missingness of that measurement is clearly not *missing completely at random* (MCAR). However, as the selection depends only on the observed data (X_0, Y_1), the missing data are *missing at random* (MAR). Handling missing data plays an important role in the analysis, and for that there are several methods, which can be used. In this section we present a likelihood-based approach and Section 5.2 considers a multiple imputation approach for carrying out the analysis.

Next, we will use the following indexing of individuals. Individuals $j = 1, \ldots, n$ are measured both at the baseline and at the halfway of the follow-up and individuals $j = n+1, \ldots, N$ have the baseline measurement only. We also divide individuals without second measurement into two groups so that individuals $j = n+1, \ldots, n'$ have not had an event before the time of the second measurement and thus could have been selected and individuals $j = n'+1, \ldots, N$ have already had an event and were not candidates for that measurement.

The analysis can be carried out with the likelihood-based approach for incomplete data.[20] This requires also the specification of the distributions of the covariate, namely $p(x_0)$ and $p(x_1|x_0, y_1)$. The parameters associated only with the covariate process are denoted by ψ. Utilizing the assumption $p(y_2|x_0, x_1, y_1) = p(y_2|x_1, y_1)$, the likelihood becomes

$$
\begin{aligned}
L(\theta,\psi) &= p(x_0,x_1,y_1,y_2) \\
&= p(x_0,y_1)p(x_1|x_0,y_1)p(y_2|x_1,y_1) \\
&= p(x_0,y_1)\prod_{j=1}^{n} p(x_{1j}|x_{0j},y_{1j})p(y_{2j}|x_{1j},y_{1j}) \\
&\quad \times \prod_{j=n+1}^{N} \int_{-\infty}^{\infty} p(x_{1j}|x_{0j},y_{1j})p(y_{2j}|x_{1j},y_{1j})dx_{1j} \\
&= \prod_{j=1}^{n} p(x_{0j})p(t_{1j},\delta_{1j}=0|x_{0j})p(x_{1j}|x_{0j},t_{1j},\delta_{1j}=0)p(y_{2j}|x_{1j},t_{1j},\delta_{1j}=0) \\
&\quad \times \prod_{j=n+1}^{n'} p(x_{0j})p(t_{1j},\delta_{1j}=0|x_{0j}) \int_{-\infty}^{\infty} p(x_{1j}|x_{0j},t_{1j},\delta_{1j}=0)p(y_{2j}|x_{1j},t_{1j},\delta_{1j}=0)dx_{1j} \\
&\quad \times \prod_{j=n'+1}^{N} p(x_{0j})p(t_{1j},\delta_{1j}=1|x_{0j}).
\end{aligned}
$$

Missing data are treated here as integrals with respect to the missing variable over the support of it. In simple settings with only one covariate, the integrals in the likelihood function can be calculated by numerical integration, which we apply in this paper. If the covariates are categorical variables, the integration would simplify to summation and direct numerical maximization would be feasible and straightforward. In the general setting, methods such as EM-algorithm[21] or Bayesian data augmentation[22] could be used.

5.2 Multiple imputation

Another method we apply in handling missing data is multiple imputation.[23] Now, we do not have to model the marginal distribution $p(x_{0j})$ as in the previous approach, but only the distribution $p(x_{1j}|x_{0j},t_{2j},\delta_{2j})$ (the imputation model) has to be specified. It is known that when the missing data are in a covariate of the analysis model, the outcome variable should be used in imputation model.[24] In our case, this means that in addition to baseline covariate measurement, the survival information must also be used to predict the missing covariate value of the second measurement.

We use a linear imputation model of the form proposed by White and Royston[25]

$$x_{1j} = \beta_0 + \beta_1 x_{0j} + \beta_2 \delta_{2j} + \beta_3 H_0(t_{2j}) + \varepsilon_j, \qquad (2)$$

where $\varepsilon_j \sim N(0,\sigma^2)$ and the survival information is included in the predictors using the status indicator δ_2 and the cumulative baseline hazard function

$H_0(t_2)$. For normally distributed x_1 this imputation model is approximately valid when covariate effects and cumulative incidence are small.[25] White and Royston also discuss the estimation of $H_0(t)$ in the case of the semiparametric Cox model.[25] We assume instead a Weibull proportional hazards model, when we have

$$H_0(t) = \left(\frac{t}{b}\right)^a .$$

The multiple imputation, using the above imputation model, is carried out in a standard way.[23]

6 Simulation study

6.1 Description of the simulation study

Simulation studies were carried out in different settings to explore the performance of the proposed selection methods, namely D_β-optimality and D-optimality. Simulated data were made to resemble our real data of Section 7 apart from a few exceptions.

We considered a follow-up setting of 20 years with 1500 individuals. A time-varying piecewise constant covariate had the baseline measurement x_0 in the beginning and second measurement x_1 in the half-way of the follow-up. First, the baseline values x_0 of the covariate were generated from the normal distribution with mean $\mu_0 = 0$ and variance $\sigma^2 = 0.02$ and the values of second measurement were drawn from the normal distribution conditioned on the baseline measurement with mean $\mu_1 = 0.5x_0$ and variance $\sigma_\varepsilon^2 = 0.015$, which leads to the correlation of 0.5 between x_0 and x_1. Then, the ages of individuals were generated from the uniform distribution. We created datasets with a narrow age range from 35 to 44 years and a wide age range from 25 to 64 at the baseline to see whether this has an effect on the results. The age and the covariate of an individual were assumed to be independent.

Survival times were simulated from the Weibull distribution depending on the time-varying values of the covariate through the proportional hazards model with regression coefficient $\beta = 3$. We chose to use such a large value of β compared to real data, so that the possible effect of covariate stands out in the selection. The parameters of the Weibull distribution were set to shape $a = 5.7$ and scale $b = 28000$ (time scale in days), which roughly equal those estimated from the real data. The survival times of individuals who were alive at the end of the follow-up were censored. The follow-up was conducted at the same calendar time with all individuals, but at each measurement time the ages are not the same for the individuals.

With both narrow and wide age range, 1000 datasets were generated. The numbers of events were on average 101 and 187 during the first half of the follow-up and 222 and 288 during the second half of the follow-up, in datasets with narrow and wide age range, respectively.

6.2 Selection of individuals

The selection of individuals for the second measurement was carried out after ten years of follow-up by simple random sampling (SRS), D_β-optimality selection and D-optimality selection. Two different sizes, $n = 600$ and $n = 200$, of subsets selected for the second measurement were considered. In the case where $n = 600$, we selected on average 43% with narrow age range and 46% with wide age range of the individuals, who were alive at that moment. The corresponding numbers in the case where $n = 200$ are 14% and 15%, respectively.

Before presenting the results obtained with different selection methods, we examine what kind of individuals are selected according to D_β-criterion and D-criterion, in other words, who are most important to remeasure. Figure 1 shows the selection order for any n up to 600, when the selection is based on the D_β-optimality. The order is, of course, irrelevant when we analyze the data. The selection seems to prefer extreme baseline covariate values, but also the age of an individual has an effect. With the wide age range, the age is the more important factor in the selection than with the narrow age range.

From Figure 2 we see that using the D-optimality leads to rather different kind of selection compared to D_β-optimality. With the narrow age range, during the first few dozens of selection rounds, the D-optimal selection is focused on individuals with high covariate values and high age. However, the age has a greater effect on selection than with D_β-optimality, especially when we consider the wide age range.

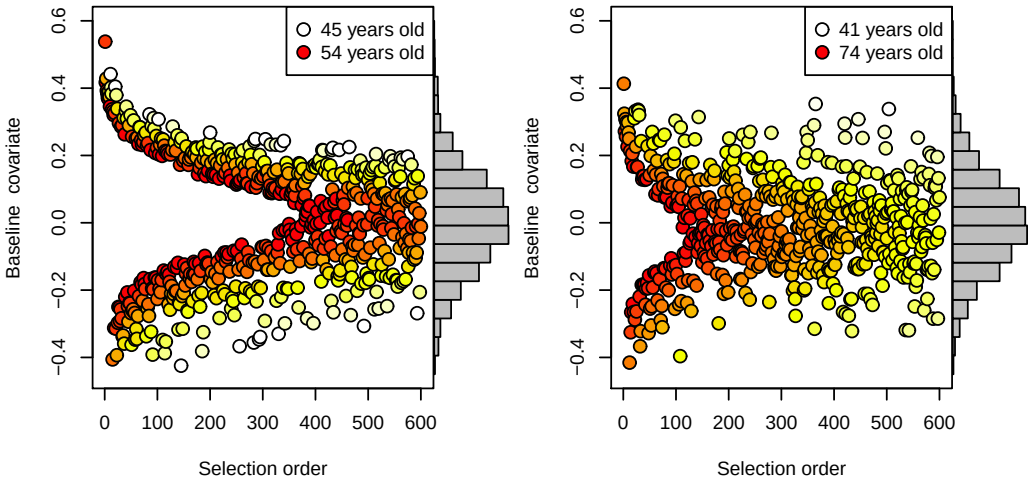

Figure 1: Selection order of individuals to be remeasured for any n up to 600 using D_β-optimality and simulated data. The left panel shows the order for data the with narrow age range and the right panel is for the wide age range. Each point corresponds to one individual: the brightness (color in the online version) shows the age of the individual at the time of the selection, the vertical axis shows the value of the baseline covariate and the horizontal axis shows the round when the individual was selected in the greedy algorithm. The histograms on the right vertical axes show the distribution of the covariate in the entire cohort. With the wide age range, the minimum age selected is 41 years although the minimum age in the cohort is 35 years.

6.3 Analyses with different designs

The aim of the selections illustrated in Figures 1 and 2 is to find the subset which would lead to as reliable as possible estimation of parameter β in our Weibull proportional hazards model. All analyses were carried out with the R statistical software.[26] With multiple imputation the `weibreg` function from the `eha` package[27] was used to fit the Weibull model. When applying the likelihood-based approach with numerical integration, the likelihood was optimized using the `optim` function with the BFGS algorithm and the standard errors were evaluated using the `hessian` function from the `numDeriv` package[28]. Integrals were calculated numerically with the `integrate` function. In Weibull proportional hazards models, especially with small samples, the profile likelihood could be considered instead of the maximum likelihood, which we are using, to improve the estimation.[29]

Estimation results using numerical integration in handling missing second measurement data are presented in Table 1. Bias seems to be negligible

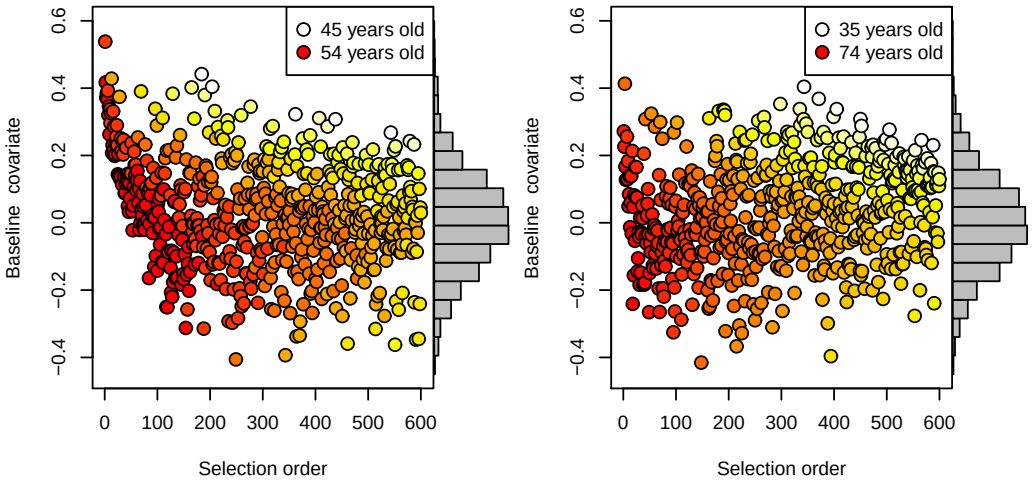

Figure 2: Selection order of individuals to be remeasured for any n up to 600 using D-optimality and simulated data. The left panel shows the order for data with the narrow age range and the right panel is for the wide age range. Each point corresponds to one individual: the brightness (color in the online version) shows the age of the individual at the time of the selection, the vertical axis shows the value of the baseline covariate and the horizontal axis shows the round when the individual was selected in the greedy algorithm. The histograms on the right vertical axes show the distribution of the covariate in the entire cohort.

in all the designs. From the standard errors, obtained from the inverse of the numerically differentiated Hessian matrix, and standard deviations of estimates, it can be seen that with the optimal selections the estimation is more precise than with the simple random sampling (SRS). The differences in the standard errors and standard deviations between the full cohort and the designs with $n = 600$, are smaller than one would expect simply considering the difference in the number of observations. Although the D_β-optimality is defined to provide the minimal variance for $\hat{\beta}$ there are no clear differences between the standard errors of the D_β-optimal and D-optimal designs.

Table 2 shows the results when multiple imputation is used in dealing with missing data. We see that there is virtually no bias. Again we see that although only a subset is selected for the second measurement, results are rather good compared to the case where everyone is measured twice.

In Tables 1 and 2 the standard errors and deviations seem to be quite consistent when $n = 600$, especially with the wide age range. Nevertheless, when we compare the results of multiple imputation and numerical integration with

Table 1: Simulation results for different designs using **numerical integration**. Bias is the mean bias of estimates $\hat{\beta}$, SD is the standard deviation of estimates $\hat{\beta}$ and Mdn(SE) is the median of the estimated standard errors of $\hat{\beta}$ from 1000 simulation runs for both narrow and wide age ranges.

	Design	Baseline ages 35–44			Baseline age 25–64		
		Bias	SD	Mdn(SE)	Bias	SD	Mdn(SE)
	Full cohort	0.01	0.40	0.40	0.01	0.35	0.33
$n = 600$	SRS	0.01	0.49	0.48	0.00	0.41	0.39
	D_β-optimal	0.01	0.46	0.46	0.01	0.38	0.36
	D-optimal	0.00	0.47	0.46	0.01	0.38	0.36
$n = 200$	SRS	0.00	0.57	0.55	0.00	0.46	0.44
	D_β-optimal	−0.01	0.56	0.53	0.00	0.42	0.40
	D-optimal	0.01	0.55	0.53	0.00	0.42	0.40

narrow age range in the case where $n = 200$, we see that multiple imputation produces clearly greater standard errors and deviations. Furthermore, in this setting the D-optimality seems to perform better than D_β-optimality. This is, however, a problem related to multiple imputation, since theoretically the D_β-optimal design cannot give larger standard error for $\hat{\beta}$ than the D-optimal design. The assumptions of the imputation model (2) do not hold, because the covariate effect or cumulative incidence cannot be considered small. As a summary of the simulation study we can say that optimal selections improve the estimation and that the two design criteria seem to lead virtually to same improvement.

7 Results for the East-West study

Data from the Finnish cohorts of the Seven Countries Study are used. The Seven Countries Study was initiated in the late 1950s to study variation in cardiovascular disease and related risk factors levels.[13] The Finnish cohorts ($N = 1711$) included all men born between 1900 and 1919 in two geographically defined areas, one in Eastern and the other in South-Western Finland, from which comes the name East-West study. The baseline survey was conducted in 1959 and re-examinations in 1964, 1969, 1974, 1984, 1989 , 1994 and 1999.[30,31] The cohorts were followed-up for mortality until the end of 2010. In these analyses data on re-examination from 1964 and 1974, and information on the age of death are used.

Table 2: Simulation results for different designs using **multiple imputation**. Bias is the mean bias of estimates $\hat{\beta}$, SD is the standard deviation of estimates $\hat{\beta}$ and Mdn(SE) is the median of the estimated standard errors of $\hat{\beta}$ from 1000 simulation runs for both narrow and wide age ranges.

	Design	Baseline ages 35–44			Baseline age 25–64		
		Bias	SD	Mdn(SE)	Bias	SD	Mdn(SE)
	Full cohort	0.00	0.41	0.40	0.02	0.34	0.33
$n = 600$	SRS	0.00	0.53	0.52	−0.02	0.42	0.41
	D_β-optimal	0.00	0.50	0.49	0.05	0.37	0.36
	D-optimal	−0.01	0.49	0.48	0.04	0.37	0.36
$n = 200$	SRS	−0.03	0.82	0.79	−0.07	0.60	0.61
	D_β-optimal	0.01	0.77	0.74	0.04	0.53	0.50
	D-optimal	-0.09	0.70	0.71	0.01	0.52	0.50

We use only a part of the data so that we consider the measurement of the year 1964 as the baseline measurement, 1974 as the year of the second measurement and 1984 as the end of the follow-up, when censoring is carried out. As a time-varying covariate, we use diastolic blood pressure, which is logarithmic and centered in the analyses. Survival time is considered to be the age at the time of death. After removing individuals who were already dead before the measurement of the year 1964 and who do not have observations of diastolic blood pressure, we have 1501 eligible individuals for our study. In this setting, we have 354 events in the first half and 424 events in the second half of the follow-up.

The selection order with D_β-optimality and D-optimality can be seen in Figure 3. D_β-optimality seems again to prefer extreme covariate values and at a fixed value of covariate, older individuals are selected first. In D-optimality, the age is clearly more important factor in the selection. All in all, the selection with the East-West data looks very similar compared to one with the simulated data, which could have been expected, because the simulated data were made to resemble these real data.

It turned out that the numerical integration did not work well with the real data. Different initial values of the optimization function led to different estimates, which may have arisen from invalid distributional assumptions. Thus, only the estimation results using multiple imputation for missing data are presented. Table 3 shows that there are some differences in the estimates, especially with D-optimal design when $n = 200$. Relative to the standard

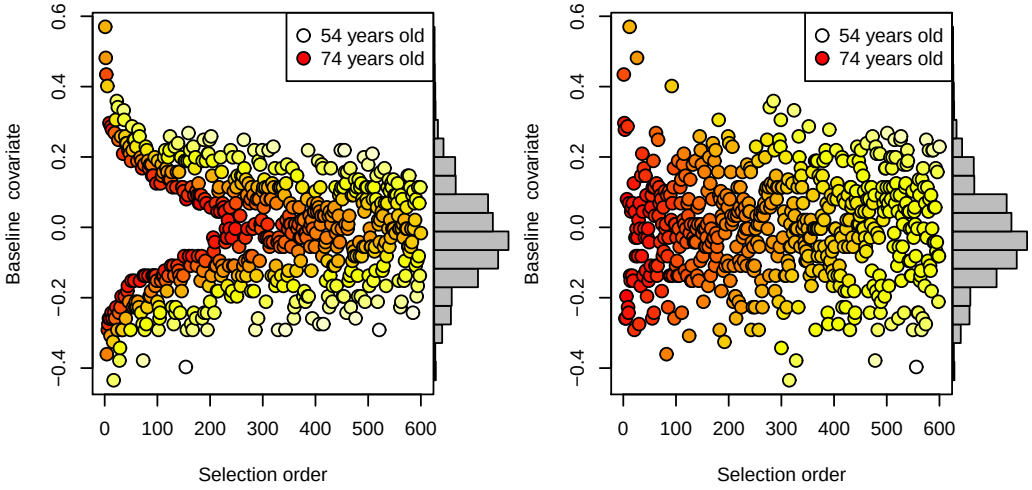

Figure 3: Selection order of individuals to be remeasured for any n up to 600 in the East-West data using D_β-optimality (left panel) and D-optimality (right panel). Each point corresponds to one individual: the brightness (color in the online version) shows the age of the individual at the time of the selection, the vertical axis shows the value of the baseline covariate and the horizontal axis shows the round when the individual was selected in the greedy algorithm. The histograms on the right vertical axes show the distribution of the covariate in the entire cohort.

errors, the D_β-optimal seems to be the best. By fitting the model separately for younger and older age groups, we see that the same model does not hold for younger and older individuals. This explains the problems with the D-optimal design, since it consists mainly of individuals with high age.

It is worth noticing, that the gain from the optimal designs is free of charge once the selection procedure has been implemented. Comparing the designs with $n = 600$, we approximate that reducing the standard error of the SRS-design to the level of the D_β-optimal design would require about 189 more individuals to be remeasured. That is, the efficiency of the SRS-design is 75% in this comparison. This kind of amount of measurements would create notable additional costs in epidemiological studies, like the East-West study. In the case where $n = 200$, approximately 23 more individuals are required in the SRS-design to reach the standard error of the D_β-design, which means that the SRS has here the efficiency of 90%.

Table 3: Results with the East-West data for different designs using multiple imputation. For the simple random sample (SRS) Estimate is the median of estimates $\hat{\beta}$ and SE is the median of the estimated standard errors of $\hat{\beta}$ from 100 analyses with random sample for the second measurement.

	Design	Estimate	SE	Individuals required to reach the SE of the D_β-design
	Full cohort	0.91	0.27	
$n = 600$	SRS	0.91	0.32	189 (32%) more
	D_β-optimal	0.89	0.30	
	D-optimal	0.71	0.30	
$n = 200$	SRS	0.97	0.45	23 (12%) more
	D_β-optimal	0.85	0.44	
	D-optimal	0.53	0.48	

8 Conclusion

Limited resources lead us to investigate the cost-effectiveness of study designs. In this paper, we considered the selection of individuals for the often costly re-examination of a time-varying covariate in a follow-up study. Two different Fisher information based optimality criteria, D_β-optimality and D-optimality, were applied and compared to the SRS using simulated data and real epidemiological follow-up data.

The selections carried out according to the optimality criteria indicate that individuals with extreme baseline covariate values and high age would be most important to remeasure. The criteria balance differently between these characteristics: D_β-optimality stresses the extremity, whereas D-optimality prefers old individuals. With both criteria, age is the more important factor in the selection when the age range of individuals is wide than when the age range is narrow. Results from the analyses with different designs show that, when handling missing data with multiple imputation or numerical integration, the precision is usually better for the optimal selections than for the SRS. No clear differences between the two optimal selections were observed. Numerical integration looks better than multiple imputation according to the simulation results, but from the real data we have learned that it may be sensitive to the model assumptions.

When the proportion of the missing data is large, the estimation may be sensitive to the model misspecification.[32] However, it should be emphasized

that the model used for the optimal selection may be different from the model used in the analysis. After the second measurement it is possible to check the validity of the distributional assumptions and change the model if needed.

The future work will consider more complicated designs where several repeated measurements will be carried out on several covariates. The current work forms a basis for these extensions.

Acknowledgements

Authors are grateful to an anonymous referee for valuable comments. We also thank Jukka Nyblom, Harri Högmander, Jouni Helske and Satu Helske for helpful comments and suggestions.

Funding

The research of the first author was supported by the Emil Aaltonen Foundation.

References

1. Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, et al. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. Am J Epidemiol. 1999;150(4):341–353.

2. Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable. J R Stat Soc Ser A Stat Soc. 2000;163(2):173–189.

3. Reilly M. Optimal sampling strategies for two-stage studies. Am J Epidemiol. 1996;143(1):92–100.

4. Liu X. Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. J Educ Behav Stat. 2003;28(3):231–248.

5. Atkinson AC, Donev AN, Tobias RD. Optimum experimental designs, with SAS. Oxford: Oxford University Press; 2007.

6. Pukelsheim F. Optimal Design of Experiments. New York: Wiley; 1993.

7. Karvanen J, Kulathinal S, Gasbarra D. Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. Comput Stat Data Anal. 2009;53(5):1782–1793.

8. Mehtälä J, Auranen K, Kulathinal S. Optimal designs for epidemiologic longitudinal studies with binary outcomes. Stat Methods in Med Res. December, 13, 2011;DOI: 10.1177/0962280211430663.

9. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53(1):330–339.

10. Diggle PJ, Sousa I, Chetwynd AG. Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. Stat Med. 2008;27(16):2981–2998.

11. Rathbun SL, Song X, Neustifter B, Shiffman S. Survival analysis with time varying covariates measured at random times by design. J R Stat Soc Ser C Appl Stat. 2013;62(3):419–434.

12. Njagi EN, Molenberghs G, Rizopoulos D, Verbeke G, Kenward MG, Dendale P, et al. A flexible joint modeling framework for longitudinal and time-to-event data with overdispersion. Stat Methods in Med Res. July, 18, 2013;DOI: 10.1177/0962280213495994.

13. Keys A. Coronary heart disease in seven countries. Circulation. 1970;41(1):186–195.

14. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.

15. Robert CP, Casella G. Monte Carlo statistical methods. New York: Springer; 1999.

16. Elfving G. Optimum Allocation in Linear Regression Theory. Ann Math Stat. 1952;23(2):255–262.

17. Wright SE, Bailer AJ. Optimal experimental design for a nonlinear response in enviromental toxicology. Biometrics. 2006;62:886–892.

18. Dykstra O. The augmentation of experimental data to maximize $|X'X|$. Technometrics. 1971;13:682–688.

19. Retout S, Comets E, Samson A, Mentré F. Design in nonlinear mixed effects models: optimization using the Fedorov–Wynn algorithm and power of the Wald test for binary covariates. Stat Med. 2007;26(28):5162–5179.

20. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.

21. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of Royal Statistical Society Series B. 1977;39:1–38.

22. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. J Am Stat Assoc. 1987;82(398):528–540.

23. Rubin DB. Multiple imputation for nonresponse in surveys. Wiley Ser Probab Math Statist. New York: John Wiley & Sons, Inc.; 1987.

24. Moons KG, Donders RA, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–1101.

25. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28(15):1982–1998.

26. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2012. ISBN 3-900051-07-0. Available from: `http://www.R-project.org/`.

27. Broström G. eha: Event History Analysis; 2012. R package version 2.2-0. Available from: `http://CRAN.R-project.org/package=eha`.

28. Gilbert P, Varadhan R. numDeriv: Accurate Numerical Derivatives; 2012. R package version 2012.9-1. Available from: `http://CRAN.R-project.org/package=numDeriv`.

29. Ferreira da Silva M, Ferrari SL, Cribari-Neto F. Improved likelihood inference for the shape parameter in Weibull regression. J Stat Comput Sim. 2008;78(9):789–811.

30. Karvonen MJ, Blomqvist G, Kallio V, Orma E, Punsar S, Rautaharju P, et al. Men in rural East and West Finland. Acta Med Scand. 1966;180(s460):169–190.

31. Pekkanen J. Coronary heart disease during a 25-year follow-up: risk factors and their secular trends in the Finnish cohorts of the Seven Countries Study. PhD Thesis. University of Helsinki. Government Printing Centre; 1987.

32. Saarela O, Kulathinal S, Karvanen J. Secondary Analysis under Cohort Sampling Designs Using Conditional Likelihood. J Probability Stat. 2012;Article ID 931416, 37 pages. doi:10.1155/2012/931416.